Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3291–3296
Marseille, 11–16 May 2020

A Representative Corpus of the Romanian Language: Resources in Underrepresented Languages

Ludmila Midrigan - Ciochina, Victoria Boyd, Lucila Sanchez Ortega, Diana Malancea-Malac, Doina Midrigan, David P. Corina
University of California, Davis,
Diana.malancea@gmail.com, doinamidrigan@gmail.com
{lmidrigan, vaboyd, lsan, dpcorina} @ ucdavis.edu

Abstract

The effort in the field of Linguistics to develop theories that aim to explain language-dependent effects on language processing is greatly facilitated by the availability of reliable resources representing different languages. This project presents a detailed description of the process of creating a large and representative corpus in Romanian – a relatively under-resourced language with unique structural and typological characteristics, that can be used as a reliable language resource for linguistic studies. The decisions that have guided the construction of the corpus, including the type of corpus, its size and component resource files are discussed. Issues related to data collection, data organization and storage, as well as characteristics of the data included in the corpus are described. Currently, the corpus has approximately 5,500,000 tokens originating from written text and 100,000 tokens of spoken language. it includes language samples that represent a wide variety of registers (i.e. written language - 16 registers and 5 registers of spoken language), as well as different authors and speakers

Keywords: Romanian, corpus representativeness, corpus design

Introduction

As a field, theoretical linguistics has moved from idealized descriptive accounts of language toward more explanatory theories that strive to ascertain the effects of language experience on language processing and human cognition in general. Along with this shift has come a greater appreciation of the need to consider linguistic phenomena in all the world's languages. This effort has been greatly aided by the development of computational approaches and digital resources in many of the world's languages. The availability of reliable, easy to access resources is crucial for different linguistics subfields. Bender (2009, 2016) argues for the importance of "linguistic knowledge" referring to ways in which different languages vary in their structure, for the development of language-independent natural language processing systems. Corpora are an essential tool in studying typological patterns across and within languages, documenting and preserving natural and endangered languages and developing computational tools for language processing (Gippert, Himmelmann and Mosel, 2006; Crystal 2000; Bradley and Bradley 2002; etc.). Hence, the project described in this paper aims at creating a such resource in one of the world's languages that remains relatively underrepresented and understudied to date. This paper provides a detailed description of the process of creating a representative corpus in Romanian, an Eastern – Romance language with unique grammatical and typological characteristics. The corpus is not only a large but also a balanced repository of written and spoken samples of language in Romanian, that can be used as a reliable tool in linguistic studies.

Historically, Romanian has lacked large collections of empirical linguistic data, which has made it difficult across the decades to provide solid, empirically motivated analysis and study of the Romanian language. Although generally characterized as a Romance language, Romanian has its unique grammatical characteristics, representing a test case for those interested in less classically analytic languages. Unlike other Romance languages, Romanian has kept its strong usage of Latin case-marking and a rich declension system (Kihm, 2012) while developing typological characteristics it shares with Balkan rather than Romance languages as an effect of language contact (Hill, 2004; D'hulst, Coene and Avram, 2004). Presently, it is spoken by approximately 24 – 26 million people as a native language and about 4 million people as a secondary language. As a function of its unique grammatical and typological characteristics, Romanian is of interest to linguists. In recent years, new resources have been developed, that enable and facilitate the study of the language. Among these are the Romanian treebank corpora, included in the Universal Dependencies (UD) Project, that contains the following typology of texts: The Romanian Non-standard UD treebank, called UAIC-RoDia (Maranduc, 2017) with approximately 16,190 sentences; the SiMoNERo (Mitrofan et al., 2019), which is medical corpus of contemporary Romanian extracted from the Biomedical Gold Standard Corpus for the Romanian Language (BioRo) (Mitrofan and Tufiș, 2018), as well as The Romanian UD treebank, called RoRefTrees (Mititelu, Ion, Simionescu and Irimia, 2016), containing 9500 trees annotated according to Universal Dependencies. Other resources, not included in the UD project, are the CHILDES database (Child Language Data Exchange) (MacWhinney, 1996), containing three small Romanian Corpora that represent child language and roTenTen16 (Kilgarriff, 2014), which is a web-based corpus. Though a very large resource, the web-based corpus data is not well balanced (Kilgarriff, 2007) (i.e. web-based language samples are not intentionally selected to proportionall represent different registers[1] of the language, different authors, specific time period, etc.). The Moldavian and

[1] The term *register* is used in literature as an umbrella term referring to general or more specific language varieties defined by situational characteristics (i.e. language used in novels, vs journal articles, vs conversations (Biber, 1995), or more specific: language used in the novels of Victor Hugo, language used in the writings of Shakespeare, etc.).

Romanian Dialectal Corpus (MOROCO), that contains over 10 million tokens collected from the news domain and representing the Romanian dialect spoken in Romania as well as the Republic of Moldova (Butnaru and Ionescu, 2019). The largest, to date corpus of the Romanian language is the Reference Corpus of the Contemporary Romanian Language (CoRoLa), 1,257,752,812 tokens. The data in the corpus is distributed in an unbalanced way, containing language samples from the legal, administrative, scientific, journalistic, imaginative, memoirs and blogposts domains. The motivation to create a new resource for Romanian was to build a balanced repository of the language that includes as many registers as possible, written as well as spoken, from different Romanian authors, regardless the spoken dialect. We envision to add as many tokens of spoken data as we will collect for written data.

1. Building of The Romanian Corpus

2.1 Planning the Building of The Corpus.

Corpora designers need to carefully address the following general practices when creating a corpus: the planning of the corpus construction, including decisions concerning the corpus type, size, representativeness and balance; data collection and storage, including obtaining copyright permission, creating the metadata and cleaning the text; and finally decisions referring to corpus annotation (Biber, Conrad & Reppen, 1998; McEnery and Wilson, 1996; Mayer, 2002). Based on these practices, our project follows the following methodological decisions, made prior to data collection:

1. The size of the corpus will reach at least 4 million words.
2. The corpus will contain at least 15 different registers.
3. Each register will contain approximately 100,000 words.
4. We will try to control for variables such as gender, age of the speaker or writer.
5. Individual text files will be saved in UTF-8 format and stored in individual directories, hierarchically representing all registers.
6. Information on a variety of variables such as author names and gender, type of texts (i.e. full versus shorter samples), and the online source of the texts will be stored for further reference.
7. The corpus will be a balanced monitor[2] corpus.
8. Future steps: adding written text from earlier time periods (1500 – 1800), adding spoken language samples and annotating the corpus for different grammatical markers.
9. These decisions have guided the data collection process. In the following sections, some of these considerations are discussed.

2.2 Corpus Type.

In order to enlarge the scope of our resource, we created a *monitor* corpus. A corpus that allows constant additions of new samples of data not only increases its size and representativeness as access to new data is gained, but also represents language through time and at its current state. A corpus as such can be used for a wide variety of linguistic measures, but also for typological studies and lexicography. Text is continuously being added to this resource as well as spoken language and transcribed spoken text.

2.3 Corpus Size and Representativeness.

Various factors can influence the ability to collect language data (e.g. time, data availability, funding, etc.). Classically, a *representative* corpus will include natural language samples that represent as many instances of language usage as possible. "Lengthier corpora are better than shorter corpora. However, even more important than the sheer length of the corpus is the range of genres included within it" (Meyer, 2002); thus, we aimed at including a wide range of genres from various language domains. Another aspect of the language we tried to include is the dialectal varieties in both written and spoken language (i.e. Moldovan Romanian). Although the spoken dialects of the Romanian spoken in the two countries (i.e. Romania and Republic of Moldova) differs due to the strong Russian influence on the spoken language in Moldova (Baar and Jakubek, 2017), the literary standard is similar (Minahan, 2013). Both dialects are included in the corpus.

The initial goal for this project was to collect a minimum of 4 million words. This goal was attained (i.e. at its current state, the corpus has approximately 5,500,000 tokens from written data and approximately 100,000 tokens of spoken data); however, samples will continually be added. It is worth mentioning that depending on its purpose, some authors argue for small rather than large corpora. For example, O'Keeffe, McCarthy & Carter (2007), argues for the concept of small corpora as means to encourage detailed analysis of each individual feature; however, a multi-purpose corpus requires larger and more representative samples of language data (i.e. quantitative measures such as word frequency, neighborhood density, affix productivity, etc. seem to benefit from larger data samples). Biber (1990) found that many grammatical features are well stable within 1,000-word samples; however, rarer grammatical features may still be underrepresented in such small samples. The Balanced Romanian Corpus (BRC) has a collection of full texts for a wide range of registers. All registers contain over 100,000 words (see Table 1). Some registers are represented in a larger proportion for both opportunistic and intentional reasons. We were able to include certain texts over others since we obtained permission from a limited number of sources, which contain only specific language genres. We also included lengthier samples in certain genres, depending on the genre characteristics. For example, novels tend to be longer than poems, since we decided to include full texts for all genres, it was necessary to allow a larger proportion of tokens in order to include larger number of writings. The larger register in the corpus is *Literatura Tradusa* 'Translated Literature'. This was done intentionally, with the aim to mirror the genres included in the BRC written originally in Romanian. As these works were originally written by non-Romanian authors, the original language may have influenced the translations and we wanted to represent these peculiarities. The translated literature contains: Eseuri ('Essays'), Fabule ('Fables'), Fictiune ('Fiction'), Filosofie ('Philosophy'), Poezii ('Poems'), Romane ('Novels') and Teatru ('Theater'). The large number of tokens in the register of translated

[2] Monitor corpora do not have a fixed size, reflecting ongoing changes in the language, as data continues to be added.

literature was a consequence of trying to add literature originally written within different genres and in different languages (i.e. English, French, Spanish and Russian; we plan to add texts from other languages as well).

We tried to equally represent text samples from both male and female authors. However, some registers (i.e. *Romane,* 'Novels', *Romane Istorice*, 'Historical Novels' and other registers) have predominantly male authors in Romanian literature; thus, balancing the genders represented was challenging. Table 1. below shows the number of authors in each register with some of their demographic characteristics. The steps taken while collecting and editing the data were documented for further reference. Also, a list of the specific web pages and the names of the authors related to each document was separately created and stored. The process of getting the copyright permission was also documented.

Written Data			
Research Articles	42	60	102
Fairy Tales	2	0	2
Fiction	4	1	5
Fiction-Romance	3	0	3
Philosophy	31	10	41
History	5	2	7
Translated Lit.	24	9	33
Textbooks	11	18	29
Memoirs	1	0	1
Christian Poems'	48	38	86
Poems	18	10	28
Novels	6	0	6
Mystery Novels	3	0	3
Historical Novels	4	0	4
News	18	9	27
Theater	4	0	4
Total Number of Authors	**224**	**157**	**381**
Percentage of Total	**58.79%**	**41.21%**	**100.00%**
Spoken Data			
Interviews	2	3	5
Lifestyle Show	9	10	19
Dora's Show	4	2	6
Total Number of Authors	**15**	**15**	**30**
Percentage of Total	**50.00%**	**50.00%**	**100.00%**

Table 1. Number of Authors in the BRC, by Register.

3 Methods and Results

3.1 Corpus Data

The sources for the text in the corpus were decided based on both "judgment" (i.e. trying to create a language repository that is balanced and representative across registers) and "convenience" (i.e. different registers of the language were selected but were also restricted by copyright permissions). For the spoken data, we have so far obtained the permission of bloggers and journalists/TV hosts Veronica Ghimp-Deineco and Lilia Lozovan Roşca, as well as producer and journalist Ana Danilescu to include some of their posts and TV show series. For each of the sources (including Audio Data), written permission was obtained from either the website owner or the author/writer/speaker or the producer of the sample. Below is a list of the web resources used:

List of Sources:

Written Data

http://bsclupan.asm.md/ - Biblioteca Științifică Centrală "A. Lupan" (Central Scientific Library "A. Lupan")
www.resursecrestine.ro - "Biblioteca Resurse Crestine" ("Library of Christian Resurces")
biblioteca@upsc.md – "Biblioteca Științifică UPSC" ("Scientific Library UPSC") moldstat@statistica.gov.md – "Biroul Național de Statistică al Republicii Moldova" (National Bureau of statistics of Moldovan Republic"
http://www.bibnat.ro/ - "Biblioteca Națională a României" ("National Romanian Library")
https://www.zdg.md/ - *Ziarul de Garda* ("Warder Newspaper")
https://www.publika.md – *Știri* 'News'

Spoken Data

https://www.publika.md – *Știri* 'News'; Interviews by Veronica Ghimp – Deineco
http://www.canal2.md/category/emisiuni/vorbe-bune – *Vorbe bune cu Lilu* 'Good words with Lilu' TV Show.
https://www.jurnaltv.md/category/dora – *Dora Show* 'Dora's Show' Comedy Show. *Italia Patria Noastră* 'Italy our Home – Country'.
https://www.youtube.com/watch?v=8Yr1b9D71UM – Veronica Gimp
https://www.youtube.com/watch?v=i87xXmhiwbg – *Despre Eva* 'About Eva'

3.2 Romanian Corpus Registers

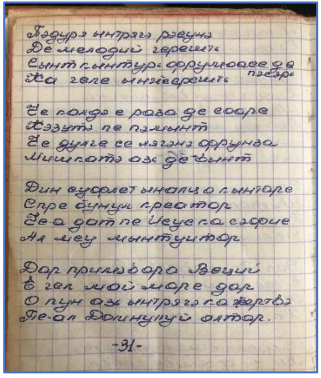
Although the BRC at its present state includes a smaller proportion of transcribed spoken language, a wide variety of registers were chosen while collecting the written text. Within the registers, different authors and rubrics were included. For example, within the register *Știri* 'News', journal articles from the rubrics *Justiție* 'Justice', *Social* 'Social', *Politic* 'Politics', *Editoriale* 'Editorials' were equivalently collected; in *Manuale* 'Textbooks', pieces of text from *Biologie* 'Biology', *Chimie* 'Chemistry', *Istorie* History, *Muzică* 'Music' as well collections of *Interviuri* 'Interviews' and *Bibliografii* 'Biographies' are proportionally represented; *Articole Cercetări* 'Research Articles' contains text from articles about sports, mathematics, physics, pedagogy, and medicine; *Poiezii* 'Poems' include poems for children, and different genres of poetry; *Basme* 'Fairy Tales', include text written in prose as well as lyrics, etc. Table 2 below gives a list of all registers in the Romanian Corpus, with their respective number of tokens and percentage of the whole corpus' tokens contained within each register.

Genre	Number of Tokens	Percentage of Total
Written Data		
Research Articles	453,117	9.22%
Fairy Tales	190,242	3.66%
Fiction	308,248	5.32%
Fiction-Romance	118,450	2.20%
Philosophy	238,200	4.59%
History	435,565	8.09%
Translated Lit.	1,794,414	33.52%
Textbooks	145,393	2.76%
Memoirs	171,316	3.21%
Christian Poems'	147,546	2.65%
Poems	107,502	1.92%
Novels	583,480	10.70%
Mystery Novels	149,311	2.76%
Historical Novels	161,763	3.06%
News	142,931	2.73%
Theater	210,998	3.61%
Total	**5,358,476**	**100.00%**
Spoken Data		
Interviews	20,346	19.09%
Lifestyle Show	76,002	71.33%
Dora's Show	10,201	9.57
Total	**106,549**	**100.00%**

Table 2. Number of tokens and types generated by NLTK from the Romanian Corpus.

The corpus includes registers that have been considered valuable for the corpus representativeness as well as for the documentation of the Romanian language at its present state. For example, while children's literature has not been largely considered in linguistics research (Knowles & Malmkajer, 1996), Baker and Freebody (1989) analyzed texts from 163 primary school readers and noted that the frequency distribution of words in these shows different patterns compared to traditional corpora based on adult language samples (e.g. the word *little* was almost as frequent as the determiners in traditional English corpora). Hence, children's literature was considered an essential register of the language and was included in the BRC. The corpus includes two genres of children's literature: 'Fairy Tales' and 'Poetry', included within *Basme* 'Fairy Tales' and *Poezii* 'Poems', respectively. Along with fairy tales, many of the children's poems are only orally transmitted through generations. These cannot be found in print therefore, collecting the available online samples was considered crucial for language preservation purposes. We also obtained two notebooks of manually written songs (for both children and adults), collected during 1940-1950, in Republic of Moldova by Olga Midrigan, transmitted from her parents and grandparents. These are written in Romanian, using the Cyrillic alphabet (see Figure 1). We are currently transcribing them using the Romanian Alphabet. We are planning to include these in the corpus for preservation purposes. About 20% of the register *Poezii* 'Poems' is composed of children's poetry. Poetry is one of the richest compartments of children literature in Romanian (Stanciu, 1968, 2000); also, children's poetry, especially *lullabies*, have influenced many musical genres (e. g. *Doina* - a free-rhythm, highly ornamented improvisational tune – their lyrics' common themes are melancholy, longing (*dor*), love for nature, complaints about life, religious, etc.,), hence have important cultural and linguistic connotations.

Another register that was considered valuable for corpus representativeness was school textbook language samples. Since most children use textbooks through their education process, concurrent with the developing of language skills, representing this genre was considered necessary. An important language characteristic used more frequently in textbooks appears to be the imperative mood: *Rețineți definiția* 'Note the definition'. Textbook language samples were thus included as a separate genre in the corpus.



Пэдуря ынтряга рэсунэ
Де мелодий черешть
Сынт кынтурь фрумоасе де пэсэрь
Ка челе ынжечерешть

Че блондэ е роза де соаре
Кэзутэ пе пэмынт
Че дулче се лягэнэ фрунза
Мишкатэ оэ де вынт

Дин суфлет ыналцэ о кынтаре
Спре бунул креатор
Че-а дат пе чеуе на сэфие
Ал меу мынтуитор

Дар примэвара вешний
Е чел май маре дар
О пун азь ынтряга ка жертвэ
Пе-ал Домнулуй алтар.

-31-

Figure 1. Olga's Notebook.

3.2 Corpus Text Structure

Constructing enormous collections of machine-readable text from online resources is fairly easy in certain languages; however, manually collecting, parsing, reformatting, and restricting text to be in line with the corpus text encoding conventions is a time-consuming process, as it was in our case. All texts were manually extracted from the online sources to ensure good data quality, to reduce the risk of including texts that lacked the proper use of Romanian diacritics (e.g. "ţ", "î", "ă", "ş", etc.), and to facilitate the recording of the metadata. Information about the name, gender, age and nationality of the author for each text for which it was available, was manually recorded and then included in the metadata files. The proofreading was also manually performed. This was necessary especially for registers such as *Teatru*, 'Theater', where names of the actors, scenes, and acts, needed to be delete, as well as for the register *Manuale* 'Textbooks', for which many rubrics in the books contained numbers, exercises, tables, and other content that may not represent language use per se, rather mathematical and statistical facts; these were manually extracted. Individual texts were converted to UTF-8 format and saved as plain text documents. Each file was named with author name and work title. Each text was organized in separate files, in distinct directories, containing the specific register and the metadata files associated with the files. These were organized in such ways for ease of compilation and accessing. The audio data was transcribed by Romanian native speakers and saved in its own directory with the associated audio files and metadata. Audio data is being continuously added to the corpus. These were then made available online through GitHub Pages. The Romanian Corpus, including transcribed spoken data, can be easily accessed and downloaded at the link: https://lmidriganciochina.github.io/romaniancorpus/.

4 Metadata

Storing information that describes the properties of the linguistic resource and variables containing information about every individual file is very essential for both accessing information of interest for specialized studies (e.g. psycholinguistic studies looking at gender differences and language use) but also indexing and searching the corpus. For the creation and storing of the metadata for the Romanian Corpus, we are using the Arbil tool, developed at Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands **(**Withers, 2009; 2012), available at (http://tla.mpi.nl/tools/tla-tools/arbil/). This tool allows the usage of standardized profiles and schemas for both spoken and written language resources, CMDI (Component MetaData Infrastructure) framework. We are continually editing these as new data is added.

5 Corpus Annotation Aims

We are in the process of starting to annotate the Corpus, using the text preprocessing module TTL (Ion, 2007), developed in Perl. The module is available at http://ws.racai.ro/ttlws.wsdl and offers a *Sentence Splitter, Tokenizer, Tagger, Lemmatizer, and a Chunker* procedures for Romanian (Tufiş, Ion, Ceauşu and Ştefãnescu, 2008). We also envision parse tree annotation, semantic labeling and affix annotation in future steps.

6 Conclusions and Further Directions

Computerized corpora in the different languages spoken around the world has important implication for linguistic theory. Although not all questions can be answered by studying language as represented in corpora, they can greatly expand our understanding of language and its complex facets. Evidence from corpora allows researchers to document natural language, study language typology and the effects of language-dependent factors on language processing. It has also important applications in the development of natural language processing systems. Thus, building resources in languages that are understudied is crucial. This project's goal was to enable linguists to study Romanian and its unique grammatical characteristics, by creating a reliable repository of text that represents a wide variety of registers of this language.

Although the corpus presented in this project is one of the largest available for Romanian, we want to continue enlarging this resource, specifically the spoken data; Compiling large samples of spoken data is not an easy task. Accurately transcribing spoken language can be time consuming and expensive; it also requires native speaker knowledge. However, spoken language is by far the mode in which language is most frequently used, and it has its distinct characteristics. In written language, the authors tend to clean the text in somewhat unnatural ways: the ideas are well formed, and well organized. While when producing language, the speakers tend to make many repetitions (i.e. same words are said twice, or even more than two times), utter unfinished thoughts (Mayer, 2002), and produce various speech errors. Spoken language tends to have generally shorter sentences, with words that may not appear at all in written language (i.e. *aha* is used a lot in conversational Romanian to show 'agreement' or 'approval', while it is not a word that appears in written text). Language dialects are yet another reason why spoken language may have different characteristics than written language. Brysbaert and New (2009), found that frequencies that are calculated from movies' subtitles and television were better than the ones found on written text. Thus, a further direction for the development of the corpus is adding and including an equal amount of spoken data along with the written samples. Another step is adding text samples representing different time periods in the history of Romanian language development (i.e. current Romanian orthography is little more than a century old (Mallinson, 1988), thus writings representing the original language forms may be found particularly in print). Further work is yet needed in order to make the resource as easy to use as possible; the current corpus is at its initial stage of annotation. Many corpora used in various linguistic studies are still unannotated; however, working with annotated corpora makes the process of information retrieval easier and faster. One further steps of the present project is annotating the text and transcribed spoken language for various types of grammatical information. Some of the target annotations are word-class annotation as well as morphological annotation, including affix annotation for different types of affixes of the language; in addition, parse tree annotation and semantic labeling are also considered.

The BRC project, aims to create a new, reliable resource in Romanian - a language that has unique structural characteristics, still remains understudied.

Acknowledgements

We want to thank Kenji Sagae for guidance through the process of constructing the corpus.

References

Baar, V., & Jakubek, D. (2017). Divided National Identity in Moldova. *Journal of Nationalism, Memory & Language Politics, 11*(1), 58-92.

Baker, C. & Freebody. P. (1989). *Children's first schoolbooks*. Oxford: Basil Blackwell.

Bender, E. M. (2009, March). Linguistically naïve! language independent: why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* (pp. 26-32).

Bender, E. M. (2016). Linguistic typology in natural language processing. *Linguistic Typology*, *20*(3), 645-660.

Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based approaches to issues in applied linguistics. *Applied linguistics*, 15(2), 169-189.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and linguistic computing*, 5(4), 257-269.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*.

Bradley, D., & Bradley, M. (Eds.). (2013). *Language endangerment and language maintenance: An active approach*. Routledge.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4), 977-990.

Butnaru, A. M., & Ionescu, R. T. (2019). Moroco: The moldavian and romanian dialectal corpus. *arXiv preprint arXiv:1901.06543*.

Crystal, D. (2000). *Language death*. Ernst Klett Sprachen.

D'hulst, Y., Coene, M., & Avram, L. (2004). Syncretic and analytic tenses in Romanian. The Balkan setting of Romance. *Olga Mišeska Tomić (éd.), Balkan Syntax and Semantics, Amsterdam/Philadelphia, John Benjamins Publishing Company*, 355-376.

Gippert, J., Himmelmann, N., & Mosel, U. (Eds.). (2006). Essentials of language documentation (Vol. 178). Walter de Gruyter.

Hill, V. (2004). On left periphery and focus. *Balkan Syntax and Semantics*, 339-354.

Joseph, J. (2004). *Language and identity: National, ethnic, religious*. Springer.

Kihm, A. (2012). Old French and Romanian Declensions from a Word and Paradigm Perspective and the Notion of "Default Syncretism". *Revue roumaine de Linguistique*, *57*(1), 3-34.

Kilgarriff, A. (2007). Googleology is bad science. *Computational linguistics*, *33*(1), 147-151.

Kilgarriff, A., P. Rychlý, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. In *Proc Eleventh EURALEX International Congress.* Lorient, France.

Knowles, M., & Malmkjær, K. (1996). Language and control in children's literature (1st ed.). London: Routledge.

Leech, G., & Smith, N. (2000). Manual to accompany the British National Corpus (Version 2) with improved word-class tagging. Lancaster, England: Lancaster University.

MacWhinney, B. (1996). The CHILDES system. *American Journal of Speech-Language Pathology*, 5(1), 5-14.

Mallinson, G (1988). Rumanian. In Harris, M., & Vincent, N. (Eds.). (1988). *The Romance languages* (pp. 391-419). London: Croom Helm.

Maranduc, C. (2017). A Diachronic Corpus for Romanian (RoDia). *Proceedings of the LT4DHCSEE in conjunction with RANLP*, 1-9.

Maranduc, C. (2017). A Multiform Balanced Dependency Treebank for Romanian. *Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP*, *17*, 9-18.

McEnery, T. and Wilson, A. 2001. *Corpus Linguistics. An Introduction.* Second edition. Edinburgh: Edinburgh University Press.

Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge University Press.

Minahan, J. (2013). *Miniature empires: a historical dictionary of the newly independent states*. Routledge.

Mititelu, V. B., Ion, R., Simionescu, R., Irimia, E., & Perez, C. A. (2016). The romanian treebank annotated according to universal dependencies. In *Proceedings of the tenth international conference on natural language processing (hrtal2016)*.

Mititelu, V. B., Tufiş, D., & Irimia, E. (2018, May). The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mitrofan, M., & Tufiş, D. (2018, May). Bioro: The biomedical corpus for the romanian language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mitrofan, M., Mititelu, V. B., & Mitrofan, G. (2019, August). Monero: a biomedical gold standard corpus for the romanian language. In *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 71-79).

O'keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Papandropoulou, I., & Sinclair, H. (1974). What is a word? Experimental study of children's ideas on grammar. *Human Development, 17*(4), 241-258.

Stanciu, I., (2000). *Calcuri populare şi calcuri savante în limba română,* 139-145.

Stanciu, I. (1968) *Literatura pentru copii*. Editura Didactica si Pedagogica, Bucuresti.

Tufiş, D., Ion, R., Ceauşu, A., & Ştefănescu, D. (2008, May). RACAI's Linguistic Web Services. In *Proceedings of the 6th Language Resources and Evaluation Conference-LREC*.

Withers, P. (2012). Metadata Management with Arbil. In V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou, & M. Monachini (Eds.), Proceedings of the workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR at LREC 2012, Istanbul, May 22nd, 2012 (pp. 72-75). European Language Resources Association (ELRA).

Withers, P. (2009). Presentation on the use of Arbil for editing metadata and archiving in the Clarin context. Talk presented at Internal meeting. Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. 2009-05-15.

Xue, N. (2003). Chinese word segmentation as character tagging. International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing, 8(1), 29-48.